# Automatic writer identification framework for online handwritten documents using character prototypes

Guo Xian Tan, Christian Viard-Gaudin, A. C. Kot

# Automatic Writer Identification Framework for Online Handwritten Documents Using Character Prototypes

*Guo Xian Tan*
Nanyang Technological University of Singapore
tanguoxian@pmail.ntu.edu.sg

*Christian Viard-Gaudin*
IRCCyN/UMR CNRS 6597 Ecole Polytechnique de l'Université de Nantes
christian.viard-gaudin@univ-nantes.fr

*Alex C. Kot*
Nanyang Technological University of Singapore
eackot@ntu.edu.sg

## Abstract

*This paper proposes an automatic text-independent writer identification framework that integrates an industrial handwriting recognition system, which is used to perform an automatic segmentation of an online handwritten document at the character level. Subsequently, a fuzzy c-means approach is adopted to estimate statistical distributions of character prototypes on an alphabet basis. These distributions model the unique handwriting styles of the writers. The proposed system attained an accuracy of 99.2% when retrieved from a database of 120 writers. The only limitation is that a minimum length of text needs to be present in the document in order for sufficient accuracy to be achieved. We have found that this minimum length of text is about 160 characters or approximately equivalent to 3 lines of text. In addition, the discriminative power of different alphabets on the accuracy is also reported.*

**Keywords**: Writer identification, information retrieval, online handwriting, fuzzy c-means, allographs

## 1 Introduction

Writer identification from handwritten documents reflects new insights into the evolving nature of handwriting research. This applied field of handwriting lies at the frontiers of handwriting research, relying on both multi-disciplinary and inter-disciplinary nature of the entire research field, spanning from psychology and behavioral sciences to pattern recognition and data mining. Writer identification systems must be clearly distinguished from writer verification systems. As defined by Srihari et al., the problem of writer verification is to make a decision as to whether two handwriting samples were written by the same writer [1]. Writer verification performs a comparison between a questioned document and one or more documents from the same known writer in order to ascertain the authenticity or identity of the questioned document. On the other hand, writer identification involves executing a search in a database of documents on a basis of a small snippet of handwriting [2] and returns a ranked list of results for the search, analogous to a search for people based on

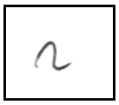
handwriting samples. The difference, though subtle, lies in the applications in which they can be utilized in. A potential application of writer identification can be found in forensic expert decision-making systems where a narrowed-down list of identified writers provided by the writer identification system can be made available to the forensic experts, where they can then further scrutinize the list and take a final decision. This kind of application aids the forensic expert to make a more informed choice, in a reduced amount of time. Another implementation lies in identifying writers for writer adaptation in mobile devices. On the contrary for writer verification, a typical application can be to integrate the writer verification system into an authentication system that can be used to monitor and regulate the access to certain confidential sites or data. The writer verification system can then make an informed decision and decide whether the person trying to gain access into the site is authorized for access or to reject the person.

Online document indexing using writer information provides three-fold distinct advantages. Firstly, from information security's point of view, writer identification has ubiquitous applications in digital rights management and forensic analysis. The individuality of handwriting styles allows handwriting analysis to be considered as a behavioral biometric trait that can differentiate between different people [3, 4]. Handwriting analysis has traditionally been used in the field of forensic document analysis by forensic experts in the detection and to a large extent, the deterrence of fraud, identity theft and embezzlement cases. Secondly, in environments where large amounts of documents, forms, notes and meeting minutes are constantly being processed and managed, knowing the identity of the writer would provide an additional value. One such application is to process and retrieve the identities of students for subsequent verification purposes. Thirdly, we can also perform writer-adaptation to create, store or retrieve a profile of handwriting styles of the writers if we are able to automatically determine their identities [5, 6]. This way, the performance of the handwriting recognition system can be vastly improved since we are able to customize the recognition system to tailor to the writing profile and style of the writer. One can thus imagine a virtuous circle where knowing the writing identity from the bootstrap recognition system will help to improve the recognition results as already mentioned above.

Handwriting recognition has over the past few decades matured to a stage where readily available commercial and industrial text recognition engines are able to provide us with reasonably high recognition accuracies [7, 8]. We envisage that future research directions in writer identification will be developed based on the foundations of existing technology in order to fully take advantage of the level of maturity and accuracies of the current handwriting recognition technology. Keeping this in mind, this paper proposes a framework for an automatic segmentation approach using an industrial text

recognition engine that splits the text into characters in order to match them with allographic[1] prototypes. Such allographic prototypes that are defined at the character level represent the individual handwriting styles of different writers. The advantage of working at the character level allows for a more consistent set of prototype templates to be defined and built [9]. The proposed framework then adopts a fuzzy c-means algorithm to create a distribution of frequency vectors that statistically models the handwriting styles of the writers. The distribution of handwriting styles then undergoes classification to identify the writer of the documents. Writer identification is then essentially accomplished by a matching process between the allographic prototypes of the writers in question to templates of allographic prototypes of the reference writers found in the database. A simplified description is presented in Table 1. There are four different prototypes of allographs for alphabet '*r*' displayed. The reference documents provided by writer *i* and writer *j* are transformed into a frequency vector based on the distribution of different styles of allographs for alphabet '*r*'. The test document from writer *T* would undergo the same transformation. Following this, distances would be computed using the test document's vector with those stored in the reference. The top-1 ranked reference writer would be identified as the writer of the test document. In the case shown in Table 1, the distribution of frequency vectors for reference writer *i* is more similar to that of test writer T and therefore writer *i* is identified as the same writer as test writer T.

**Table 1.** Simplified example of the proposed methodology.

| | | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|
| Available prototypes for character 'r' (selected by k-means algorithm) | |  |  |  |  |
| Reference Documents | frequency vector of writer *i* | 0.71 | 0.00 | 0.00 | 0.29 |
| | frequency vector of writer *j* | 0.10 | 0.00 | 0.60 | 0.30 |
| Test Document | frequency vector of writer *T* | 0.70 | 0.20 | 0.00 | 0.10 |

The remainder of this paper is organized as follows: Section 2 provides a brief survey of some of the main works in online writer identification to review the current state-of-the-art in this domain. Section 3 then describes the proposed framework and experimental setup. Preprocessing such as normalization and resampling and our proposed fuzzy c-means

[1] Allographs are different shapes and forms of the same alphabet.

algorithm is discussed in this section. Section 4 then presents the experimental results. Finally, discussions and future directions to explore for online writer identification are given in section 5.

## 2 Previous works

Much progress has been made in the field of writer identification in the last decade. Writer recognition systems can typically make use of global features such as texture, curvature and slant features [10-12] as well as a combination of local features such as graphemes, allographs and connected components [1, 13, 14] to identify the writers. They can be generally classified into approaches that utilize text-dependent or text-independent techniques. Signatures are examples of text-dependent systems since the writers have to write the exact same text as what they have written previously for the system during the enrolment process. Srihari et al.'s [3, 4] works falls into this category of text-dependent approaches. They proposed the use of two levels of features; one at the macro level, making use of features such as the average slant, aspect ratios and entropies at the paragraph or document level. The other level functions at the micro level and makes use of features such as gradient, structural and concavity at the word or character level. They then used a multi-layer perceptron for writer verification and obtained an accuracy of 98% with this text-dependent approach that only required limited amount of text to be present.

Our work falls into the latter category of text-independent techniques where the writers are not bounded by any specific lines of text in order for the system to recognize them. Instead, the system analyzes their handwriting styles through a series of automated processes, regardless of what they have written. This kind of writer recognition systems included previous works such as the method proposed by Pitak et al. [15] which adopted a Fourier transformation approach. The extracted features are the velocities of the barycenter of the pen movements and they are transformed into the frequency domain using Fourier transform. The advantage in adopting such a model is that it is text-independent, but at the expense of a lower noise tolerance. The noise must be filtered out as much as possible in the pre-processing stage, otherwise the noise might be mistaken for high velocity components once the features are transformed into the frequency domain. Text-independent writer recognition systems can also make use of stochastic approaches like the Hidden Markov Models (HMM) technique presented in the works of Schlapbach et al. [16]. They built one HMM model for each writer and extracted nine features at the line level from a database of 100 writers. An identification rate of 96% was attained based on 8600 text lines from the 100 writers.

Of paramount importance to this paper are the approaches that make use of graphemes or allographs, which has been gaining popularity in writer identification. Such state-of-the-art algorithms and techniques make use of a template matching approach that assigns handwriting styles to prototype templates which are representative of the handwriting styles [9, 17-20]. The prototypes attempt to model the writing styles of the writers as close as possible, based on features extracted from the online handwritten documents. Identification of the writer is then achieved based on the comparison of some similarity measures between the extracted features from the reference writers and the test writer in question. Bensefia et al. [19, 20] proposed using a sequential clustering approach at the grapheme level for offline texts to categorize different writers for their writer identification system. This approach attained an identification rate of 86% on an English database of 150 writers and 95% on a French database of 88 writers. The advantage of this method is that it does not depend on any lexicon and is therefore language independent. However, working at the grapheme level does not ensure a high level of consistency in reproducing the set of templates for writer identification. Another work that involves graphemes is by Bulacu et al. [11]. They have generated a codebook of graphemes and combined them with features at the texture level to attain an identification rate of 92% on a data set of 650 writers. Both of these works make use of information retrieval techniques in their systems.

Information retrieval (IR) techniques are gaining popularity in writer identification. Among many popular types of IR models such as the fuzzy model, Boolean model or the probabilistic model [21, 22], the vector space model approach that was first proposed by Salton et al. [23] remains to this day one of the most dominant approaches in IR due to its relatively simple, yet effective design. This vector space model involves two stages; an indexing phase and a retrieval phase in a high dimensionality feature space. The indexing phase involves representing the set of documents with a set of occurrence vectors, the term frequency (tf) and inverse document frequency (idf). The underlying principle of the tf-idf combination relies on how frequent (or infrequent for the case of idf) a feature occurs in the document to represent the relevance of that feature towards the retrieval of the document. Consequently, the retrieval phase then compares the tf-idf vector of the query document with that of the indexed document for retrieval of the document. This IR approach was later adapted for use by Bensefia et al. [19] in the context of writer identification. Their contribution was to apply the concepts of tf and idf using graphemes as their feature set for the problem of writer identification. They proposed that the invariant handwriting styles of writers can be viewed as features that are defined within the writer's set of allographs. A clustering algorithm can then be used to define the groups of patterns that model the invariant handwriting styles of writers. Subsequently, a tf-idf score will be calculated based on these prototypic groups of patterns for the indexing and retrieval phase. The works by

Bensefia et al. laid down the foundations of numerous current writer identification techniques that adopt the IR approach. Among recent research in writer identification using IR techniques lies two noteworthy works by Niels et al. [17, 18] and Chan et al. [9] that yielded promising results.

Niels et al. [18] used dynamic time warping to hierarchically cluster allographs and build a set of membership vectors, which contains the frequency of occurrence of each allograph for each character. This prototypic template of membership vectors then represented the handwriting styles of the different writers. However, dynamic time warping approaches are computational expensive. Furthermore, it is difficult to cover all variations in handwriting during the training of the prototypes and dynamic time warping is highly sensitive to the absence of prototypes. Therefore, dynamic time warping will not be able to give the expected results if a new variation in handwriting that was not previously covered during the training of the prototype were to appear [18]. Another distinctive difference is that the framework proposed in this paper is fully automatic whereas the work done by Niels et al. relies on a manual segmentation process. The work by Chan et al. [9] handled this issue of previously missing handwriting variations during prototype building by adopting a statistical approach. They made use of a character prototype distribution to model the specific allographs used by a given writer and created statistical distributions to model the handwriting styles of writers. A top-1 accuracy of 95% was achieved based on this text-independent approach which considered 82 reference writers. Even though working at the character level as opposed to using the grapheme or word level appears to be quite challenging, character based approaches are able to produce a more consistent set of templates for writer identification provided that recognition and segmentation are performed accurately.

From the review of the current state-of-the-art described in this section, the following conclusions can be derived. Firstly, the text-dependent approaches allow high accuracies to be achieved even from a limited sample of text. However, one serious drawback of this is the issue of feasibility in implementing this kind of systems in reality. Writers will have to know the exact text to write, thus restricting its applicability to limited real-life situations. Text-independent approaches, on the other hand, circumvent this issue by using statistical methods that extract writer-specific features that are insensitive to the textual content of the documents. The drawback is that a minimum amount of text needs to be present for such methods to be statistically sufficient. Since our proposed methodology falls into the latter category, we have conducted a study to determine the minimum amount of text required, which is presented in section 4.5. Secondly, the literature review allows us to conclude that prototype-matching based approaches are gaining popularity in recent years, as they are able to

provide high levels of accuracies. Table 2 provides a brief summary of some of the recent works in writer identification for the past decade.

**Table 2.** A brief survey of writer identification for the past decade.

| Author | Year | Approach | Accuracy | Language | Domain |
|---|---|---|---|---|---|
| Zois et al. [24] | 2000 | Morphological approach | 96.5% on 50 writers<br>97% on 50 writers | English<br>Greek | Offline |
| Said et al. [25] | 2000 | Gabor filters & grayscale co-occurrence approach | 95% on 20 writers | English | Offline |
| Srihari et al. [3] | 2002 | Text-dependent multi-layer perceptron approach | 98% on 1000 writers | English | Offline |
| Pitak et al. [15] | 2004 | Fourier transformation approach | 98.5% on 81 writers | Thai | Online |
| Schlapbach et al. [16] | 2004 | Hidden markov models approach | 96% on 100 writers | English | Offline |
| Bensefia et al. [26] | 2005 | Grapheme-based clustering approach | 95% on 88 writers<br>86% on 150 writers | French<br>English | Offline |
| Bulacu et al. [11] | 2007 | Textural and allograph prototype approach | 92% on 650 writers | English | Offline |
| Neils et al. [17] | 2008 | Allograph prototype matching approach | 100% on 43 writers | English | Online |
| Chan et al. [9] | 2008 | Discrete Character prototype distribution approach | 95% on 82 writers | French | Online |
| Our approach | 2008 | Continuous Character prototype distribution approach | 99% on 120 writers | French | Online |

Interestingly, we observe a trend of more and more IR-based approaches [9, 11, 17, 26] being proposed for writer identification in recent years. This gain in popularity is proof of the potential that such IR models can achieve for writer identification, in spite of the simplicity in its design. This upsurge in popularity can first be credited to early Bensefia et al.'s works [19] for adapting IR models into writer identification. Our proposed methodology draws inspiration from Bensefia et al.'s works. However, our approach presented here differs on several notable exceptions: (1) Our methodology proposes a prototyping approach at the character level, instead of the grapheme level. The advantage of working at the character level is that robust and consistent prototypes can be utilized. Furthermore, prototypes that are built at the character level allows for an intuitive graphical inspection of the prototypes that can be invaluable for helping forensic experts to analyze handwriting. In addition, our character prototyping approach is not limited to just the Roman script, but the approach can be extended to other scripts that make use of a set of alphabets for their writing systems such as the Greek and Cyrillic scripts. (2) This paper provides a design for a fully automatic framework that integrates an industrial engine into the writer identification system. The significant benefit from using an industrial engine lies in being able to

exploit current advancements in handwriting technology. (3) The fuzzy c-means algorithm proposed in this paper has allowed remarkable improvements in accuracies to be attained.

## 3 Proposed Writer Identification Framework

One of the key originalities of the proposed writer identification framework is the usage of an approach that performs an automatic segmentation and labeling of the text at the character level. Handwriting recognition has in recent years, reached a level of maturity where readily available commercial and industrial text recognition engines are able to provide us with reasonably high recognition accuracies [8]. An industrial character segmentation and recognition engine, "MyScript SDK" [27], with the French linguistic resource attached for increased accuracy, has been used for this purpose. It is out of the scope of this paper to describe the core of this recognition engine in details as we have simply used it as an off-the-shelf product to integrate into our framework. The proposed framework can be divided into three stages, namely the prototype building stage, the reference and test document indexing stage and finally, the retrieval stage [28, 29], as illustrated in figure 1.

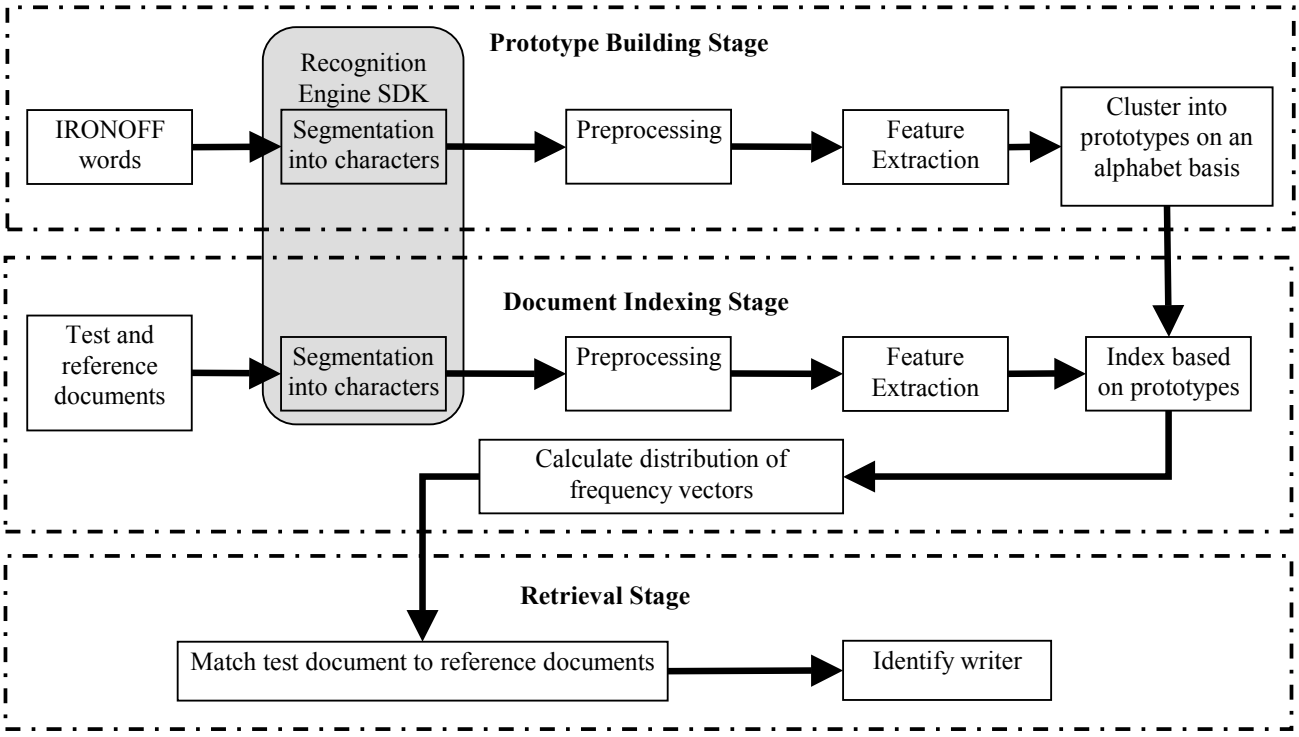
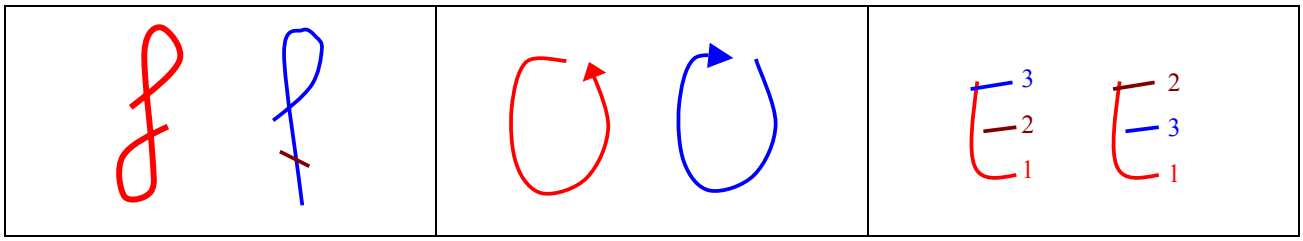



**Figure 1**. Block diagram for proposed methodology.

### 3.1 Prototype Building Stage

During the prototype building stage, prototypes are clustered at the character level, using the IRONOFF database [30] of 16585 isolated French words that are written by 373 subjects. The purpose of this stage is to build a set of character prototypes by using characters in context of words to model the different allographs of the 26 Latin alphabets (‘a’ to ‘z’). The industrial text engine automatically segments the isolated words from the IRONOFF database into a total of 87719 characters, after which the twenty-six respective subsets that have been obtained are used to build allographic prototypes that model the handwriting styles of different writers. Allographic prototypes at the character level can exploit three different types of handwriting variations to perform writer identification, specifically (1) morphological variations, illustrated in Table 1 with the example of alphabet ‘r’ and in figure 2a, (2) directional variations and (3) temporal variations, as illustrated in figures 2b and 2c. We therefore make use of the natural existence of such diversities in handwriting to differentiate between different styles of writing during the clustering of our prototypes and in the subsequent stage of document indexing..




(a) (b) (c)

**Figure 2(a)**. Morphological variation. **Figure 2(b)**. Directional variation. **Figure 2(c)**. Temporal variation.

#### 3.1.1 Clustering into Character Prototypes

After the characters are segmented, the segmented characters then underwent further preprocessing where the size of each segmented character is normalized and resampled to a fixed number of 30 points [31, 32]. A process of feature extraction on each of the resampled points is then carried out. The features being used are the x and y co-ordinates (2 features), the directions of the local tangent expressed by cosine and sine of the angle (2 features), the local curvatures (2 features) and the binary Pen-up or Pen-down information (1 feature). This effectively allows us to work using a feature space of dimension: 30 points × 7 features = 210. The extracted set of features derived from the IRONOFF isolated words database are then clustered into representative character prototypes, using the well-established k-means clustering

algorithm [33]. Figure 3 illustrates some character prototypes of the alphabet 'f' obtained after clustering. The dotted lines represent the trajectory when the pen is in the Pen-up position. The k-means clustering algorithm is performed on an alphabet basis, thereby giving us 26 alphabets × N prototypes. We have chosen N=10 experimentally, which is discussed in details in section 4.4.

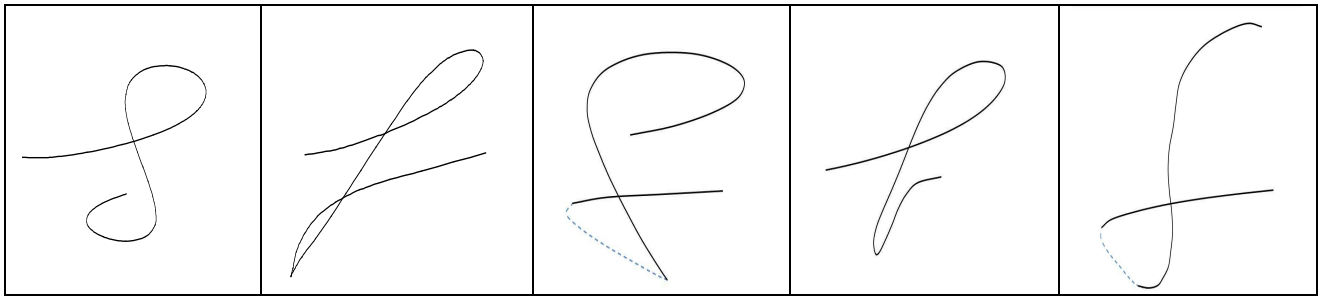


**Figure 3.** Examples of character prototypes of the alphabet 'f' after clustering.

## 3.2 Document Indexing Stage

In the document indexing stage, the sets of reference and test documents are automatically segmented and recognized into characters by the industrial engine; similar to what was undertaken in the prototype building stage. More details on the data set used are given in section 4.1. The main purpose of this indexing stage is to represent the handwriting style of each writer with a statistical distribution of term frequencies and inverse document frequencies used in the IR model [23] by mapping the features in the documents to the set of prototypes built previously during the earlier prototype building stage. This mapping of the segmented and recognized characters into a statistical distribution of prototype frequencies is accomplished by a fuzzy c-means algorithm.

### 3.2.1 Fuzzy C-means Algorithm

The previous prototype building stage serves to identify common individual handwriting traits and styles to create individual prototypes at the character level. Subsequently, the document indexing stage now utilizes these prototypes to estimate individual statistical distributions of handwriting styles for each of the test and reference documents in the database. Based on the results of the distributions, they provide statistical information about the handwriting styles of each writer. Our proposed method adopts a fuzzy c-means algorithm [33] which uses a kernel function to estimate the probability that a character $x$ has been generated by a prototype $p$. Three different kernels are proposed and described in Eq. 1 to Eq. 3. The first one is an exponential kernel with a tuning parameter to adjust the selectivity of the exponentials,

the second one is the Gaussian kernel and the third one is the inverse kernel [34]. The results of the different kernel designs are presented in section 4.3.

$$P_\alpha(p_k \mid x) = \frac{\exp\left(-\beta \times dist(p_k, x)\right)}{\sum_{k'=1}^{N} \exp\left(-\beta \times dist(p_{k'}, x)\right)} \tag{1}$$

$$P_\alpha(p_k \mid x) = \frac{\exp\left(-dist^2(p_k, x)\right)}{\sum_{k'=1}^{N} \exp\left(-dist^2(p_{k'}, x)\right)} \tag{2}$$

$$P_\alpha(p_k \mid x) = \frac{\frac{1}{dist(p_k, x)}}{\sum_{k'=1}^{N} \frac{1}{dist(p_{k'}, x)}} \tag{3}$$

$P_\alpha(p_k \mid x)$ is the probability that a given segmented character *x,* which has been recognized as of the alphabet $\alpha$, $\alpha \in$ *{'a','b', ..., 'z'}*, is assigned to prototype $p_k$ , $k \in [1, N]$. This represents the partial membership of prototype $p_k$ as discussed earlier. The function $dist(p_k, x)$ represents the Mahalanobis distance. In Eq. 1, $\beta$ is a tuning parameter which has been experimentally set to be 0.01 and *N* is the number of prototypes used. Eq.2 describes the Gaussian kernel function where the distribution of the feature vectors was assumed to be Gaussian with zero mean and unit variance. Eq.3 describes the formulation for the inverse kernel function where the notations have the same meaning as in Eq. 1.

Characters from the reference and test documents are then assigned a partial membership to the prototypes based on their distance metric to the prototypes. Therefore, characters which lie further away from certain prototypes are assigned a lesser degree to that particular prototype. $P_\alpha(p_k \mid x)$ is then used to calculate on an alphabet basis, the distribution of frequency vectors; the term frequency (tf) as described in Eq. 4 and the inverse document frequency (idf) as described in Eq.5, to be used during classification. Hence $tf_{\alpha,k}$ , $idf_{\alpha,k}$ and $P_\alpha(p_k \mid x)$ are all calculated on an alphabet basis. In Eq. 4, M is the number of characters corresponding to the alphabet $\alpha$. In Eq. 5, R is the number of reference writers and ε is a small value to prevent any numerical problems.

$$tf_{\alpha,k} = \frac{1}{M} \sum_{m=1}^{M} P_\alpha(p_k \mid x_m) \tag{4}$$

$$idf_{\alpha,k} = \log \frac{\sum_{k'=1}^{N} \sum_{i=1}^{R} tf_{\alpha,k',i} + \varepsilon}{\sum_{i=1}^{R} tf_{\alpha,k,i} + \varepsilon} \tag{5}$$

It can be seen that with our fuzzy c-means approach, a character is not labeled to one particular prototype but rather, each character is distributed over every prototype of the corresponding alphabet. The experimental results obtained serves as strong evidence to attest that our proposed fuzzy c-means approach for assigning prototypes during the document indexing stage yielded a higher level of accuracy. Finally, in the last stage of retrieval, the frequency vectors are used for classification in order to identify the writer corresponding to the test document.

### 3.3 Retrieval Stage

In this last stage of retrieval, writer identification is achieved by comparing the distribution of tf-idf vector of the test document in query with the distribution of tf-idf vectors of the reference documents. We have adopted the minimum distance classifier for this purpose to classify and rank the writers according to their similarity with the test document in query, where three different metrics have been studied; namely the Euclidean distance, the Kullback-Leibler (KL) divergence [35] and a Chi-square based metric [13].

## 4 Results

### 4.1 Experimental Data Sets

Online handwritten documents were collected from 120 writers, where each writer wrote two documents: one is considered as a reference document and the other one is taken as a test document. The contents of the two documents are different where the length of the reference and test documents varies from 86 characters to 972 characters. Each writer has to copy a given text passage taken from a variety of sources such as literacy works, financial news and short notices. In addition to copying from a given text, the writer has to provide his/her own text. This allows a large variety of content to exist in the database, which does not impose any constraints on the dependency of the domain. These reference and test documents that were collected belong to a separate dataset from the IRONOFF dataset. The rationale for this is that the IRONOFF dataset is primarily used to build the set of allographic prototypes during the previous prototype building stage. Furthermore, a separate database needed to be collected because the IRONOFF database contains only isolated words and hence are not representative of actual online documents. Therefore as a consequence, the prototype set is generic and independent with respect to the actual reference database of documents from which the writer is to be identified from. The advantage of building the prototypes from an independent database is that the prototypes need only be trained once, thus

making it much more robust and scalable to be deployed across a large number of systems. Figure 4 illustrates a sample of a text passage found in a reference document and figure 5 depicts a word that has been recognized by the industrial engine, which will also then automatically provide the segmentation into characters. However, the segmentation and recognition capabilities are not perfect. There exist instances when the recognition is correct, yet the segmentation has been performed wrongly. For example, as illustrated in figure 6b, the word “économie” might have been recognized correctly, but due to the presence of the accent found in ‘é’, the industrial text engine has wrongly segmented the following character, ‘c’. This creates an instance of noise in the allograph set and will distort the result of the prototype distribution computation. We have found the character recognition accuracy of the industrial text engine to be 91% on the whole sets of reference and test documents, which indicates that 9% of the characters have been assigned to the wrong alphabet set. This result was automatically computed by taking the true label of the documents and comparing them with the output of the industrial recognition engine. The wrongly labeled characters are included in the experiment because this will reflect a real-life scenario, where it is realistic to expect a certain degree of recognition errors. It is worth to note that the recognition engine is already trained and ready-to-use, hence no specific training is required on the documents.

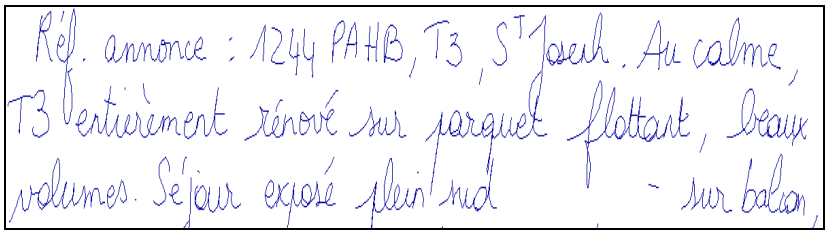
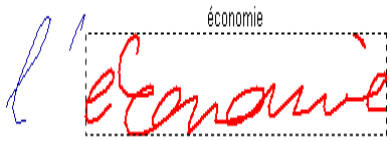


**Figure 4.** Example of a text passage from a reference document.

**Figure 5.** Example of correct segmentation and labeling at the character level by the industrial text recognizer.

a) Text to be recognized.



b) Correct recognized text “économie”by industrial text recognizer.

c) Incorrect segmentation.

**Figure 6.** Segmentation error.

## 4.2 Performance of Proposed Methodology

The first set of experiments was conducted on a set of 120 test and 120 reference documents from 120 different writers. As presented in table 3, the proposed methodology using fuzzy c-means to label the test and reference documents to the prototypes with the Euclidean distance metric resulted in a high accuracy of 98.3% for writers that are ranked correctly in the top-1 position. This translates into a misclassification error of two misclassified writers out of 120, with both of them being classified in the top-2 position. This indicates that the writer identification system has confused the misclassified documents with only one other document from a different writer. Comparisons with previous results obtained by Chan et al. [9] on the same dataset, who also performed writer identification based on character prototyping, show a significant improvement over their proposed methodology. An accuracy of 96.7% (four misclassified writers out of 120) was obtained using their method where the four writers were wrongly identified ranked at top-2, 4, 9, and 12 positions. This means that their writer identification system has confused the misclassified documents with up to 11 other documents. Therefore, our proposed methodology is able to perform with significantly higher accuracies.

**Table 3.** Performance of writer identification using different distance metrics

| Size of reference document database | **1NN[1]** | **Fuzzy C-means Approach** | | |
|---|---|---|---|---|
| | Euclidean as distance metric | Euclidean as distance metric | KL divergence as distance metric | Chi-square as distance metric |
| 120 | 96.7% | 98.3% | 91.7% | 99.2% |

This improvement over Chan et al.’s results can be explained as follows. Their methodology hinges on the concept that each character can only be assigned to one particular prototype for which a distribution of handwriting styles is built. This is flawed in reality because overlapping handwriting styles for different writers can be commonly found. Our observations reveal that there are numerous instances when the characters are close to more than one prototype in the vector space. This

can be explained by the fact that a writer can have strong, dominant handwriting style and weak handwriting styles. Weak handwriting styles change according to various circumstantial and temporal states [36, 37], which can affect the strong dominant handwriting style and lead to reminiscence of multiple overlapping handwriting styles. For such instances, the discrete allocation of prototypes used by Chan et al. does not yield good results. A writer whose dominant handwriting style does not fit well into an existing prototype will be weakly modeled using their approach. Therefore, in our proposed methodology, each character is not just assigned to one particular prototype, but rather, each character is assigned a certain degree of all the prototypes depending on how close they are to that prototype, allowing us to realize a higher accuracy. The more similar the character is to a certain prototype, the greater the degree that prototype has on the character.

$$\text{Dist}(\text{writer}_i, \text{writer}_T) = \sum_{\alpha='a'}^{'z'} \sum_{k=1}^{N} \frac{idf_{\alpha,k}\left(tf_{\alpha,k,i} - tf_{\alpha,k,T}\right)^2}{tf_{\alpha,k,i} + tf_{\alpha,k,T}} \tag{6}$$

Experiments were also performed using the Kullback-Leibler (KL) divergence and the Chi-square distance, as described in Eq. 6, as a different metric for the minimum distance classifier to determine the best performing metric for our writer identification system. We can observe from Table 3 that using KL divergence resulted in a low 91.7% identification rate. This can be attributed to the asymmetric nature of the KL divergence. It is observed that the Chi-square distance measure outperforms the Euclidean measure in our writer identification system, achieving a top-1 writer identification rate of 99.2%. This is equivalent to a misclassification error of only one misclassified writer, with the misclassified writer being in the top-2 position. The rationale, which can explain the better result obtained with the Chi-square distance, is that the Chi-square distance considers a relative difference between the two components of the distributions instead of an absolute difference as with the Euclidean distance. This relative difference is more meaningful with respect to the style of writings that we would like to distinguish. Our results also support Schomaker et al.'s results [13] that the Chi-square measure outperforms the Euclidean distance measure. Therefore, in our system, the better performing metric to use for our classifier is the Chi-square distance metric.

### 4.3 Fuzzy C-means Kernel Evaluation

Experiments were also conducted using different kernel functions [34] for the fuzzy c-means algorithm to determine the kernel function that can perform best in our writer identification system. We compared the use of three different kernel

[1] 1-Nearest Neighbor algorithm adopted by Chan et al.

functions described in Eq.1-3. Table 4 shows the comparison in performance of the writer identification among using the three different kernel functions and our results show that the exponential kernel function performs the best in our writer identification system. It can be seen that the inverse kernel function performs poorly, which can be explained by the poor behavior of such kernel functions as it approaches the centroid of the clusters. The results obtained using the Gaussian kernel function performs better than the inverse kernel function, notably because the Gaussian kernel function has the advantage of a smooth distribution as it approaches the centroid of the clusters. Our results indicate that the exponential kernel function was able to achieve the best result of 99.2% accuracy in our writer identification system.

**Table 4.** Performance of the Fuzzy c-means algorithm using different kernel functions.

| | |
|---|---|
| Identification rate using exponential kernel function (Eq. 1) | 99.2% |
| Identification rate using Gaussian kernel function (Eq. 2) | 97.5% |
| Identification rate using inverse kernel function (Eq. 3) | 96.7% |

### 4.4 Effect of Number of Character Prototypes on Accuracy

It is hypothesized that different alphabet require different number of prototypes to effectively model all the possible writing styles of that character. For example, there are more ways and styles to write the alphabet ‘f’ as compared to writing the alphabet ‘c’. For the sake of simplicity, a preliminary level of analysis has been performed to find a global optimal value for the number of prototypes needed. This analysis will allow us to better understand the behavior of the system as the number of prototypes varies. In order to verify the results of this experiment so that it can be applicable even to other databases, a cross-validation approach was used. The 120 test document database was randomly subdivided into two equal partitions of 60 test documents each, while keeping the database of 120 reference documents constant. Hence, both partitions will be making a writer identification based on the same set of 120 reference documents. The global number of prototypes is varied from 2 to 60 for every character of the alphabet.

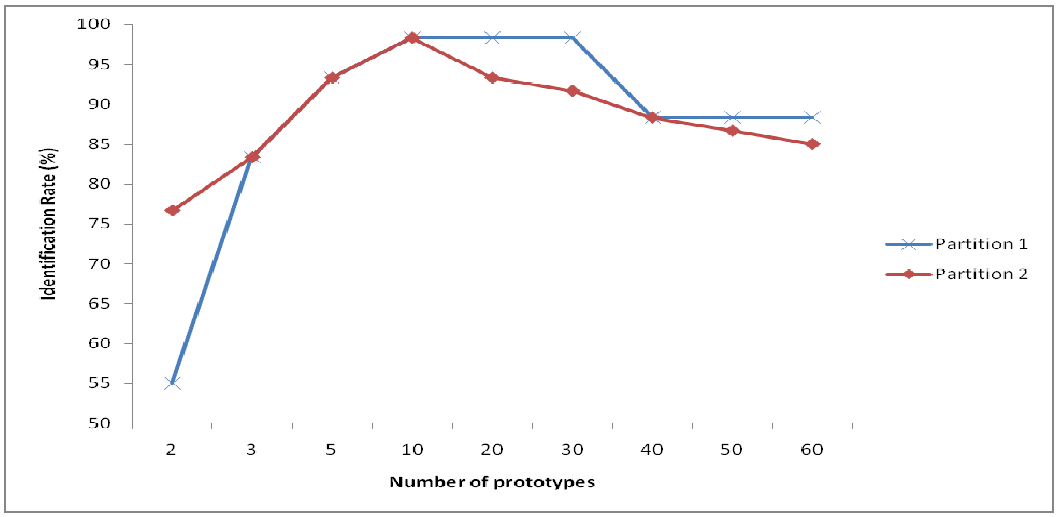



**Figure 7**. Graph of identification rate against number of prototypes.

Figure 7 shows what can be achieved by varying the number of global prototypes used for every character of the alphabet. Similar results attained by the two subsets were notably observed. As seen from figure 7, the identification rate reaches its peak when the number of prototypes used is 10. Additional number of prototypes beyond the 30 prototypes will result in a drop in the performance of the identification system. Similarly, there is degradation in the performance of the system when less than 10 prototypes are used. This can be explained by the principle of Occam's razor. A large number of prototypes create sparse dimensionality which deteriorates the performance of the classification. Likewise, insufficient number of prototypes will be unable to effectively separate between inter-class variations. Based on the above analysis, the optimum number of global prototypes is taken to be 10 in the experiment.

## 4.5 Effect of Length of Text in Test Documents

Writer identification systems that adopt stochastic approaches generally require a minimum amount of data to be present in order for the stochastic modeling to be sufficiently representative of the actual data. Therefore, it is imperative to gain an understanding of the length of text that needs to be sampled so as to facilitate the derivation of useful information that closely characterizes the handwriting styles of the writers. A series of experiments have thus been conducted to investigate the amount of text required for sufficient accuracy of the writer identification system. The experiments have been designed as follows. Only characters in the test documents have been varied, leaving the reference documents; which varies from 168 characters for the shortest reference document to 808 characters for the longest reference document, unchanged. This is because in a typical writer identification scenario, requests to identify the test

writer in question occur much more frequently than enrolment for the reference documents. Furthermore, keeping the reference documents unchanged allows a fair comparison when identifying from the same set of reference writers.

Figure 8 shows the average number of misclassified writers attained by varying the number of characters in the test documents. The desired number of characters for conducting the experiments is achieved by reducing the characters in the original test documents randomly across all the 26 alphabets. Writer identification is then performed on the test documents containing the reduced number of characters. In order to ensure that the results are more robust, this experiment is repeated over 10 different runs to obtain an average value. It is clearly shown in figure 8 that the number of misclassified writers remains approximately constant when a minimum of 160 characters are present in each test document. However, there is a severe drop in the accuracy of the writer identification system once the number of characters in each test document falls below this threshold. This can be justified by the fact that there is insufficient allographic information available to be effectively representative of the various handwriting styles. A minimum length of text is necessary to perform a reasonably accurate statistical representation of the handwriting styles. Figure 8 also highlights another interesting point where beyond this minimum threshold, any further increase in the amount of allographic information does not result in any performance enhancements in the accuracy of the system. In consequence, a minimum threshold of 160 characters or the approximate equivalence of 30 words or 3 lines in each test document is required for sufficient performance in our methodology.

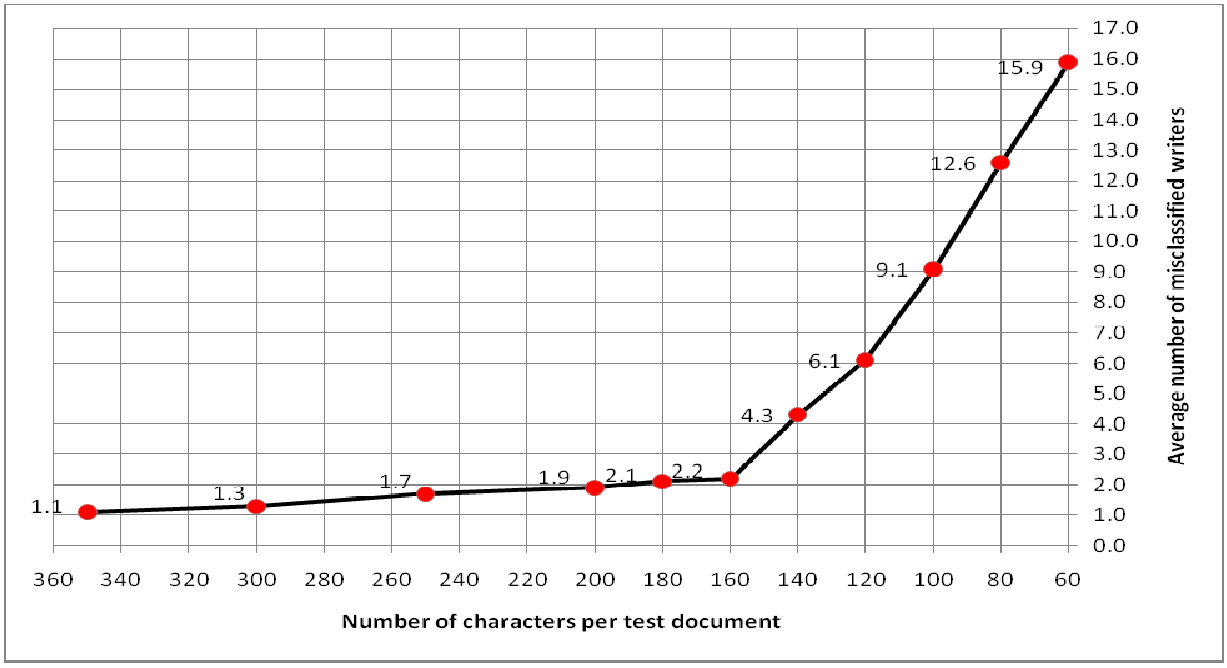



**Figure 8.** Average number of misclassifications against the number of characters in each test document (120 reference documents).

### 4.6 Discriminative Power of Alphabets

We begin by clearly defining the notion of alphabets and characters to avoid any ambiguities with the terminologies used here. An alphabet is a set of characters used in a writing system such as ‘a’, ‘b’, ‘c’ in the Latin writing system. A character is therefore an instance belonging to an alphabet. It is not unconceivable that different writers can write certain alphabets with a more distinct and unique handwriting style compared with other alphabets. This difference in handwriting styles for certain alphabets can be attributed towards multiple intrinsic and extrinsic factors. Extrinsic factors include cultural and societal influences; physical influences such as left-handedness or right-handedness, grasp and dextrality. Intrinsic variables to consider can include literacy influences and educational system influences [36, 37]. All these influences play a part to influence the writer in his handwriting style. It will then be a natural progression to hypothesize that different alphabet will have the capability to provide varying degrees of discrimination in identifying writers. For example, the alphabet ‘c’ might not have a lot of allographs and variations in its approach of writing. We thus hypothesize that such alphabets will have a low discriminative power for writer identification. In this report, we performed experiments to verify this hypothesis and showed evidence of the discriminative power of certain alphabets in writer identification.

In this experiment, only one alphabet was used at a time in writer identification. Hence, when we are investigating the discriminative power of alphabet ‘a’, only the alphabet ‘a’ was used from the reference and test documents in the document indexing stage and the retrieval stage. In this case, the size of the tf-idf vector is reduced from $26 \times N = 26 \times 10 = 260$ to only $N = 10$. The top-1 accuracy in writer identification was then obtained by considering only writers that have the alphabet ‘a’ in both their reference and test documents. Writers that do not have any alphabet ‘a’ in either the reference or test document are omitted in the ranking results. This process is then repeated for 19 more alphabets. Six alphabets, namely, ‘w’, ‘k’, ‘z’, ‘j’, ‘y’ and ‘h’ were omitted for the purpose of this experiment since these alphabets rarely appear in the documents and will skew the results if included. The outcome of this experiment is illustrated in table 5.

**Table 5.** Discriminative power of individual alphabets in writer identification.

| Alphabet | Top-1 Accuracy | % of total Characters in documents |
|---|---|---|
| a | 43.33% | 7.41% |
| s | 43.33% | 8.18% |
| d | 42.86% | 3.66% |
| t | 39.17% | 7.73% |
| r | 35.83% | 6.82% |
| e | 35.00% | 15.76% |
| o | 33.33% | 6.11% |
| i | 31.67% | 6.81% |

| | | |
|---|---|---|
| p | 30.00% | 3.04% |
| n | 27.50% | 8.11% |
| l | 23.33% | 5.23% |
| x | 20.88% | 0.53% |
| q | 19.83% | 1.14% |
| u | 19.17% | 6.68% |
| g | 18.26% | 1.38% |
| m | 16.67% | 2.89% |
| f | 14.29% | 1.01% |
| v | 13.56% | 1.61% |
| c | 12.50% | 3.26% |
| b | 11.22% | 0.98% |

From table 5, the second column indicates the top-1 accuracy obtained when only that particular alphabet is used in performing writer identification. The results supported our hypothesis that certain alphabets like ‘v’ , ‘c’ and ‘b’ might have few variations in its allographs and style of writing and hence will have a low discriminative power in writer identification. Likewise, alphabets like ‘a’, ‘s’, ‘d’ and ‘t’ are highly discriminatively in writer identification. Therefore, our results strongly indicate that different alphabets do indeed have different capabilities to provide varying degrees of discrimination in identifying writers. Some alphabets such as ‘f’ are ranked in a rather low position for its discriminative power even though we expect a high discriminative power for it. These expectations arise because having both the descender and the ascender in ‘f’ should naturally allow for more handwriting variations [33]. However, we observe that this alphabet suffers from a very low number of instances of characters: only around 1%. This leads to a poor estimation of the prototype distribution, resulting in the same problem as the experiment reported in Figure 8, i.e. when the number of characters is low the performance drops severely. Nonetheless, this experiment clearly shows that different alphabets have different identification capabilities, which supports findings from Cha et al. [4, 39]. More emphasis should be placed on such alphabets with high discriminative powers and less emphasis on those with low discriminative powers.

The third column of table 5 shows the frequency of occurrence of such alphabets in both the test and reference documents. This distribution of frequency of occurrence is similar to the results obtained by Rosenbaum et al. [40, 41] for characterizing the distribution of alphabets in general French linguistic resources, where the most frequent alphabet is ‘e’ and the least frequent alphabets are ‘w’ and ‘k’. It is interesting to note that frequently appearing alphabets such as ‘u’ (19.17% accuracy with a character frequency of 6.68%) do not provide high discriminative powers, whereas alphabets that do not appear as frequently such as ‘d’ (42.86% accuracy with a character frequency of 3.66%) is more discriminating in writer identification. This implies that the frequency of occurrence is not directly correlated to the discriminative power of

the alphabets. In our case, our *tf-idf* is not related to the frequency of occurrence of the alphabets but to the frequency of the prototypes being used, since our tf-idf has been normalized on an alphabet basis. Hence each alphabet is processed independently. Our results contrast with works by Niels et al. [17] where they conclude, without experimentation, that frequently occurring alphabets like 'e' are least suitable for distinguishing between writers and rarely occurring alphabets like 'q' are most suitable for distinguishing between writers. Based on our results, we argue that the suitability of different alphabets for writer identification should not be based solely on the frequency of occurrence of the alphabets, but rather, take into account various other intrinsic factors and extrinsic factors which can affect the discriminating power of different alphabets as well.

## 5 Discussion

From the experimental results, the proposed methodology is able to generate high accuracies of 99.2% (one misclassified writers out of 120 different writers), with this misclassified writer being identified correctly in the rank 2$^{nd}$ position. This is a remarkable improvement over Chan et al.'s [9] methodology which only attained a top-1 accuracy of 96.7% (four misclassified writers out of 120), with all the misclassified writers only being correctly identified in the rank 11$^{th}$ position. This can be explained by the fact that our proposed methodology is based on fuzzy logic that mimics the fact that a writer's handwriting style has reminiscence of multiple overlapping handwriting styles. This concludes that the system can provide high accuracies in identifying writers and is highly scalable as well. Our results also indicated that distribution matching between test and reference vectors based on Chi-square outperforms both KL divergence and Euclidean distance. This might be due to the fact that KL divergence itself is asymmetric. Furthermore, we have also showed that the optimum number of prototypes to use for our writer identification system is 10.

However, a number of considerations must be factored in. Firstly, the recognition and segmentation error of the industrial recognition engine will without a doubt influence the accuracy of the writer identification system. We argue that the current state-of-the-art industrial handwritten text recognition engines are able to obtain high levels of recognition accuracies. The industrial recognition used in this platform was able to achieve a character recognition accuracy of 91.0%, which indicates that 9% of the total number of characters has been mistakenly assigned to the wrong character prototypes. Nonetheless, the proposed methodology is still able to generate a high top-1 accuracy of 99.2%, which might be indicative of the robustness of the proposed methodology towards noise. It will be interesting to investigate, both qualitatively and quantitatively, how sensitive the proposed writer identification framework behaves to an inaccurate segmentation.

Another area of consideration to be explored from our current vantage point is to investigate how different alphabets affect the character prototypes. Certain alphabets have more writing styles than others, which suggest that more allographic prototypes might be necessary in order to fully model the entire range of writing styles for that particular alphabet. For instance, alphabets like 'f', 'h' or 'j' that have more morphological, directional and temporal variability might require more prototypes to fully represent them, as compared to other alphabets with less variability such as 'c'. In this paper, we have presented the global optimum number of prototypes to use and shown that different alphabets affect the identification rate differently. We can therefore go one step further in future to investigate a variable number of prototypes required to effectively model each alphabet, which we envisage will yield even better performance. Similarly, more research can be carried out with regards to the discriminatory power of different alphabets and the extent of its impact on the accuracy of writer identification. Our current research is spearheaded in this direction. Our eventual goal is to create a dynamic algorithm that can handle and create an alphabet matrix for each writer. The alphabet matrix will be constructed based on the discriminative power of different alphabets for each writer and will then determine if certain alphabets can be ignored for writer identification or whether more emphasis needs to be placed on certain alphabets. The main rationale for having such an alphabet matrix is to utilize only alphabets that are unique for that particular writer. This will result in a smaller feature vector set being used, which will thus reduce the computational complexity of the system.

The results clearly show that character prototype distributions can be used to effectively model handwriting styles for writer identification. It is important to note that the proposed methodology can be used to handle not just French handwritten documents, but can also include other documents that make use of the alphabet writing system, such as English, Italian or German from the same Latin script family as French. We envisage that the proposed framework can be extended to other scripts that make use of the alphabet writing system such as Cyrillic or Greek documents. It will be interesting to investigate the performance of the writer identification system where the language domain is unknown or in an environment which contains a mixture of multi-lingual documents. Our current work focuses on this aspect of developing a general framework for writer identification based on multiple languages.

## Acknowledgements

This research is jointly supported by Nanyang Technological University of Singapore, the French Merlion Scholarship and the ANR grant CIEL 06-TLOG-009.

## References


[1] H. Srinivasan, S. Kabra, H. Chen, and S. Srihari, "On Computing Strength of Evidence for Writer Verification," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, 2007, pp. 844-848.

[2] L. Schomaker, "Advances in Writer Identification and Verification," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, 2007, pp. 1268-1273.

[3] S. N. Srihari, S.-H. Cha, and S. Lee, "Establishing Handwriting Individuality Using Pattern Recognition Techniques," in *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 1195-1204.

[4] S. N. Srihari, S. H. Cha, H. Arora, and S. Lee, "Individuality of handwriting," *Journal of Forensic Sciences,* vol. 47, pp. 856-872, Jul 2002.

[5] S. D. Connell and A. K. Jain, "Writer adaptation for online handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 329-346, 2002.

[6] K. Chellapilla, P. Simard, and A. Abdulkader, "Allograph based writer adaptation for handwritten character recognition," in *10th International Workshop on Frontiers in Handwriting Recognition*, 2006, pp. 423-428.

[7] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, pp. 63-84, 2000.

[8] H. Fujisawa, "Forty years of research in character and document recognition-an industrial perspective," *Pattern Recogn.,* vol. 41, pp. 2435-2446, 2008.

[9] S. K. Chan, C. Viard-Gaudin, and Y. H. Tay, "Online writer identification using character prototypes distributions," in *Proceedings of SPIE - The International Society for Optical Engineering*, 2008.

[10] A. Busch, W. W. Boles, and S. Sridharan, "Texture for script identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, pp. 1720-1732, 2005.

[11] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, pp. 701-717, Apr 2007.

[12] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns, "Automatic script identification from document images using cluster-based templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, pp. 176-181, Feb 1997.

[13] L. Schomaker and M. Bulacu, "Automatic writer identification using connected-component contours and edge-based features of uppercase western script," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, pp. 787-798, 2004.

[14] V. Pervouchine and G. Leedham, "Extraction and analysis of forensic document examiner features used for writer identification," *Pattern Recognition,* vol. 40, pp. 1004-1013, 2007.

[15] P. Thumwarin and T. Matsuura, "On-line writer recognition for Thai based on velocity of barycenter of pen-point movement," in *Image Processing, 2004. ICIP '04. 2004 International Conference on*, 2004, pp. 889-892 Vol.2.

[16] A. Schlapbach and H. Bunke, "Using HMM based recognizers for writer identification and verification," in *Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR*, Tokyo, 2004, pp. 167-172.

[17] R. Niels, F. Gootjen, and L. Vuurpijl, "Writer Identification through Information Retrieval: The Allograph Weight Vector," in *International Conference on Frontiers in Handwriting Recognition*, 2008, pp. 481-486.

[18] R. Niels, L. Vuurpijl, and L. Schomaker, "Automatic allograph matching in forensic writer identification," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 21, pp. 61-81, Feb 2007.

[19] A. Bensefia, T. Paquet, and L. Heutte, "Information retrieval based writer identification," in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, 2003, pp. 946-950.

[20] A. Bensefia, T. Paquet, and L. Heutte, "A writer identification and verification system," *Pattern Recognition Letters,* vol. 26, pp. 2080-2092, 2005.

[21] S. M. Chen and Y. J. Horng, "Fuzzy query processing for document retrieval based on extended fuzzy concept networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 29, pp. 96-104, 1999.

[22] T. Radecki, "Generalized Boolean methods of information retrieval," *International Journal of Man-Machine Studies,* vol. 18, pp. 407-439, 1983.

[23] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM,* vol. 18, pp. 613-620, 1975.

[24] E. N. Zois and V. Anastassopoulos, "Morphological waveform coding for writer identification," *Pattern Recognition,* vol. 33, pp. 385-398, 2000.

[25] H. E. S. Said, T. N. Tan, and K. D. Baker, "Personal identification based on handwriting," *Pattern Recognition,* vol. 33, pp. 149-160, 2000.

[26] A. Bensefia, T. Paquet, and L. Heutte, "Handwritten Document Analysis for Automatic Writer Recognition," *Electronic Letters on Computer Vision and Image Analysis,* vol. 5, pp. 72-86, 2005.

[27] "Vision Objects Industrial Text Recogniser SDK MyScript Builder Help," in *SDK documentation*: http://www.visionobjects.com/about-us/download-center/_263/myscript-products-datasheets.html, 2008.

[28] G. X. Tan, C. Viard-Gaudin, and A. Kot, "Online writer identification using fuzzy c-means clustering of character prototypes," in *11th International Conference on Frontiers in Handwriting Recognition*, 2008, pp. 475-480.

[29] G. X. Tan, C. Viard-Gaudin, and A. Kot, "A stochastic Nearest Neighbor Character Prototype Approach for Online Writer Identification," in *19th International Conference on Pattern Recognition*, 2008.

[30] C. Viard-Gaudin, P. M. Lallican, S. Knerr, and P. Binter, "The ireste on/off (ironoff) dual handwriting database," *IEEE Int. Conf. Document Analysis and Recognition,* pp. 455-458, 1999.

[31] L. Vuurpijl and L. Schomaker, "Finding structure in diversity: a hierarchical clustering method for the categorization of allographs in handwriting," in *Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on*, 1997, pp. 387-393 vol.1.

[32] I. Guyon, P. Albrecht, Y. L. Cun, J. Denker, and W. Hubbard, "Design of a neural network character recognizer for a touch terminal," *Pattern Recogn.,* vol. 24, pp. 105-119, 1991.

[33] J. Han and M. Kamber, *Data Mining - Concepts and Techniques*: Elsevier, 2006.

[34] F. Hoppner, F. Klawonn, and R. K. Runkler, *Fuzzy Cluster Analysis - Methods for Classification, Data Analysis and Image Recognition*: Wiley, 1999.

[35] T. Cover and J. Thomas, *Elements of Information Theory*: Wiley, 1991.

[36] R. A. Huber and A. M. Headrick, *Handwriting Identification - Facts and Fundamentals*: CRC Press, 1999.

[37] R. N. Morris, *Forensic Handwriting Identification - Fundamentals, Concepts and Principals*: Academic Press, 2000.

[38] L. Schomaker and E. Segers, "Finding features used in the human reading of cursive handwriting," *International Journal on Document Analysis and Recognition,* vol. 2, pp. 13-18, 1999.

[39] S. Cha, S. Yoon, and C. C. Tappert, "Handwriting Copybook Style Identification for Questioned Document Examination," *Journal of Forensic Document Examiners,* pp. 1-14, 2007.

[40] R. Rosenbaum and M. Fleischmann, "Character Frequency in Multilingual Corpus 1 - Part 2," *Journal of Quantitative Linguistics,* vol. 10, pp. 1 - 39, 2003.

[41] A. A. Lyne, *The Vocabulary of French Business Correspondence: Word Frequencies, Collocations, and Problems of Lexicometric Method*: Slatkine, 1985.